Age, Sex, and Pathology Effects on Stability of Electroencephalographic Biometric Features Based on Measures of Interaction

Yvonne Höller, Arne C Bathke, and Andreas Uhl

***Abstract*—Electroencephalographic (EEG) biometric features have attracted considerable interest, but they have also drawn major criticism as having low stability. Moreover, most published studies ignore the potential effect of individual factors interacting with stability on the performance of the examined system.**

We examined the effects of age, sex, and neurological conditions in 60 subjects (a) using a single EEG recording, (b) when pooling two EEG sessions, and (c) when performing cross-session cross-validation in a biometric system based on multivariate autoregressive measures, in order to extend previous work on autoregressive coefficients.

Feature-level fusion and wrapped feature subset-selection in multivariate autoregressive coefficients (MVAR), power spectral density, and 14 additional measures derived from the MVAR model resulted in a maximum area under the curve of receiver operating characteristics (AUC of ROC)/top equal error rate (EER) of 98.85/5.34 for a single EEG, 95.51/9.70 for EEG-pooling, and 85.34/22 for cross-session cross-validation. This best result was obtained by the transfer function polynomial, which outperformed the MVAR coefficients, based on the example data set used in this study. Age, sex, and pathology significantly interacted with stability of features ($p < .001$).

We suggest further investigation of frequency-dependent measures derived from the MVAR model. We emphasize the serious problem of ignoring stability in most previously published research and recommend accurate reporting of individual factors when studying EEG biometric features in multiple sessions for enrolment and authentication on separate days.

***Index Terms*—EEG biometric features, measures of interaction, connectivity, stability**

I. Introduction

MEASURES from the brain and especially the electroencephalogram (EEG) are claimed to be good candidates for biometric systems because they are said to be universal, unique, permanent, and they can not easily be circumvented [1]–[11]. The introduction of mobile EEG sensors makes these biometric characteristics easily collectable [12], [13], so that examination of the biometric performance of characteristics derived from the EEG has attracted considerable interest in the research community.

Connectivity in the sense of statistical interdependence of human brain signals is at the forefront of neuroscientific research, which has lead to populistic names for this phenomenon, such as *connectomics* [14].

The biometric use of EEG-measures of interaction is supported by a large twin study, where especially longer connections have shown a variance that could be explained by genetic factors [15]. Coherence combined with a phase locking index and frequency characteristics yielded promising equal error rates (EER) as low as 1.63% [16]. The logical next step was the introduction of directed measures of interaction, modelling the information flow among brain regions in the sense of Granger causality [17] by autoregressive methods [18]. Most studies considered autoregression of single channels, thus ignoring the multivariate aspect of the EEG. The number and positioning of sensors are important aspects for EEG biometric systems [19]. The multivariate autoregressive model addresses the information flow between the sensors, as well as problems of volume conduction that are omnipresent in EEG recordings. This can be understood as considering the common sources of brain activity as a partial correlation between multiple signals. From this model, the so called directed measures of interaction can be derived. They are hypothesized to represent the information flow among brain regions, thus giving a more realistic model of brain activity. A study using multivariate autoregressive models reported equal error rates around 5% [20]. However, despite the promising results based on autoregressive or multivariate measures, these are not trustworthy since collection of biometric samples was based on one single EEG session with 4 participants.

As for other biometric characteristics, the EEG changes over time; on a large time scale, growth of children can affect the genetic determination of measures of interdependence [21]. In contrast, regarding a smaller time scale, some authors claim that permanence of the EEG as a biometric characteristic is high [22], [23] while other authors have stated that the major weakness of the EEG is the low reproducibility of derived biometric features [24]–[26].

It is therefore of utmost importance to assess test-retest reliability of biometric features derived from the EEG [11], [27], but most EEG studies used only one single EEG recording that was segmented into several epochs, so that enrolment and authentication tests were based on different epoch subsets, but still taken from the same EEG session [13], [16], [20], [28]–[36]. Another set of studies has included data from multiple sessions, but the data was pooled across sessions so

Manuscript submitted 28th June 2017. The presented research was funded by the Austrian Science Fund (FWF): KLIF 12-B00, I 2697-N31, and by the PMU FFF: A-11/02/004-TRI.

Y. Höller was with the Department of Neurology, Christian Doppler Medical Centre and Centre for Cognitive Neuroscience, Paracelsus Medical University; Ignaz-Harrer-Strasse 79, 5020 Salzburg, Austria (e-mail: yvonne@unak.is)

A.C. Bathke is with the Department of Mathematics, Paris Lodron University, Hellbrunner Strasse 34, 5020 Salzburg, Austria (e-mail: arne.bathke@sbg.ac.at).

A. Uhl is with the Department of Computer Sciences, Paris Lodron University, Jacob Haringer Strasse 2, 5020 Salzburg, Austria (uhl@cosy.sbg.ac.at).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2854728, IEEE Transactions on Information Forensics and Security

that training and testing sets included epochs from multiple sessions and thus, they were not disjoint in terms of sessions [22], [37]–[39]. Further methodological weaknesses of these studies were detailed in a comprehensive overview of the most recent study on stability by Maiorana et al. [40]. The natural situation would be that enrolment is done with an EEG recorded on one day, while authentication takes place on a different day [8], [40]. The EEG is very susceptible to factors like time of the day, vigilance, age, consumption of caffeine or tobacco, drugs, medication, quality of sleep, just to name a few [41]. The performance of a biometric system based on the EEG should thus be evaluated by *cross-session cross-validation*.

Table I gives an overview on previous research endeavours.

It was demonstrated that the identification rate may drop from 99% in one single session to 80.8% when two sessions were pooled, and to 46.24% when one session was used for enrolment and the other one for authentication [45]. A study with EEG recordings of 10 subjects from different days reported a classification rate of 81% [38], but in cross-session validation, the results also depended on the authentication scenario and the sample size [46]. In one study, four EEG sessions separated by a variable number of days were used, enrolment was based on three EEG sessions, and authentication was done with the fourth EEG session [22]. EER was 3.4%, but the cross-validation procedure and feature/classifier subset selection procedure was not based on three separate sets for training, evaluation, and testing. Another study [5] made use of two EEG recordings separated by several days. One EEG served for enrolment, the second one for authentication. This study reported a correct recognition rate of 98.33% on a small sample of 10 selected subjects (out of 17), where the authors commented themselves that the accuracy went down when additional subjects were introduced to the sample.

Most studies rely on the resting state, which is a state of mind during which thoughts can hardly be controlled. It has been suggested that EEG signals acquired during tasks are more stable, such as provocation of visually evoked potentials [30], [47], [48] or cognitive stimulation [49]–[52], or even multiple tasks [29], [44]. However, only a few studies so far have implemented cross-session cross-validation with cognitive stimulation (see Table I). One study was done on a very small sample of subjects (N=9), reporting an equal error rate of 7% [25]. A similar value was obtained in a study relying on ERPs [23]. In this study, the classification rate was obtained from a larger sample (N=45) in a single session, and smaller samples for a second (N=15) and third (N=9) session. Contrary to the expectation that the classification rate decreases when multiple sessions are used, classification rates were comparable, around 93%. However, the smaller sample sizes for the follow-up sessions might represent a considerable source of bias.

The most reliable results can most likely be found in the recent study by Maiorana et al. [40] alongside with another study from the same research group which uses the same sample but data acquired under cognitive stimulation [43]. The results were comparable, so that we cannot conclude that cognitive stimulation may lead to larger accuracy of the system. Most interestingly, the study by Das et al. [43] also showed that the obtained accuracy depended largely on the combination of sessions used, emphasizing session-to-session variability of task-related biometric characteristics. As a conclusion from these studies, the autoregressive components seem to be among the most promising biometric features to be derived from the EEG [40]. Thus, it is a valid question whether the further processing of these markers, in the sense of frequency zooming [26] or derivation of multiple measures within the framework of Granger Causality [53] could further improve the performance of the system.

Another potential problem with biometric systems based on the EEG has been ignored so far. The EEG is affected by factors pertinent to the individuals, such as age, sex, and pathology. These factors could affect the system participation ratio tremendously. The EEG is highly indicative for epilepsy [54]. All former studies were based on healthy participants, and/or did not take potential pathological factors into account. For example, in one study a sample of alcoholic subjects was included [4], but no comparison to healthy subjects was done, and the effect of multiple sessions was not investigated. Nevertheless, anomalies such as focal slowing or epileptic spikes can be found also in healthy subjects. Moreover, the demographic development in most western countries comes along with a rapidly growing incidence of dementia [55]. As for most other biometric characteristics, the EEG is heavily affected by healthy and pathological ageing [56]. In addition, sex affects the EEG [57], specifically the menstrual cycle in women [58]. It is most likely that sex interacts with stability of EEG biometric characteristics.

These aspects were not taken into consideration in previous studies (see Table I). Most of them enrolled mostly young and healthy samples, and most of the studies did not even report on the distribution of sex in their samples. It is highly likely that these factors play an important role especially when only small samples ($N < 10$) of healthy young participants are used for evaluation of the system (6 out of 10 studies in Table I).

We hypothesize that the effect of handling multiple EEG sessions might be highly different for subgroups with specific characteristics of age, sex, and neurological pathologies. Therefore, in order to appraise the performance metrics published in this field of research, one needs to consider the implementation of single vs. multi-session EEG datasets alongside the characteristics of the sample. A potential interaction between personal characteristics and the technical configuration with single or multiple sessions may significantly determine the conclusiveness of previously published research.

In this work we aim to document the effect of selected characteristics of samples on the performance of biometric systems, along with appropriate (multiple-sessions) vs. inappropriate (single session) configuration of the enrolment and authentication scenario. We base this examination on a set of features extracted from resting EEG. The innovative contributions of this paper are represented by the following hypotheses:

1) **Multivariate autoregressive model estimation improves performance of EEG-biometric systems.** We compare classical PSD and COH measures descriptively

TABLE I
STUDIES EXAMINING EEG BIOMETRIC FEATURES IN MULTIPLE SESSIONS WITH CROSS-SESSION CROSS-VALIDATION.

Study	N	age	sex	sessions	time	condition	features	top result	notes
[25]	9	n.a.	n.a.	12	3d	motor imagery word generation	PSD 8-30Hz	HTER 7.1%	across session learning improves performance
[27]	20	25-76	7:13	2	≥1y	rest	PSD	CRR 88%	
[4]	6	n.a.	n.a.	4		imagined speech	PSD from AR^2	CRR 99.8%	variance between sessions varies from 78.6%-99.8%
[26]	9	25-62	0:9	2	1y	movement vs. rest	AR^7 for 8-13 Hz	CRR 87.1%	movement>rest
[33]	9	n.a.	n.a.	2	1-3w	rest	AR^{12}	CRR 100%	no cross validation of post-hoc feature selection
[23]	9	18-23	5:4	3	1w - 6m	acronym viewing	ERPs	CRR 93%	accuracy stable across sessions
[42]	4	∅ 22.2	n.a.	2	≥1w	rest	time-frequency wavelet	CRR 92.33%	no cross validation of post-hoc feature selection
[43]	50	20-35	n.a.	3	7-34d	geometric shapes letters & numbers	bandpower	EER 9%	variance between session ranges approx. from EER 9-25
[40]	50	20-35	n.a.	3	7-34d	rest	AR^{10}, PSD, COH	IR 90.8%	AR>PSD>COH
[44]	20	18-28	14:6	2	248-516d	visual stimuli	ERPs	CRR 100%	no cross validation of post-hoc feature evaluation

N: sample size; age: age in years; sex: female to male ratio; time: time between sessions; d: days; w: weeks; m: months; y: years;
IR: identification rate; CRR: correct recognition rate; AR: autoregressive reflection components; acc: accuracy; ERPs: event related potentials
PSD: power spectral density; COH: spectral coherence;

to a set of measures of interaction derived from multivariate autoregressive model estimation.

2) **Age, sex, and pathology affect the performance of EEG-biometrics.** We examine EEG biometric performance in a sample of 60 participants in a wide age range and with an equal proportion of men and women, including subgroups of healthy participants, patients with temporal lobe epilepsy, elderly patients with subjective cognitive complaints, and with mild cognitive impairment.
3) **Age, sex, and pathology interact with stability of EEG biometrics.** We estimate the influence of the scenario of cross-validation in interaction with age, sex, and pathology: i) when only one EEG-session is used for enrolment and authentication, ii) when two sessions on two different days are pooled for enrolment and authentication, iii) when one session is used for enrolment and the other one (from a different day) for authentication, thus allowing for a strict cross-session cross-validation. The two EEG-sessions are separated by two weeks in order to ensure capturing two different phases of the menstrual cycle in women, and therefore maximizing the requirements for stability of the features across menstrual cycle phases.

II. METHODS

A. Feature extraction

We used a data sample as described in Section II-H with resting EEG recordings and 27 sensors, obtained during two sessions, separated by two weeks. In order to provide the same amount of data for each participant, EEG signals were shortened to the shortest available length across participants, which was 123 seconds. This signal was divided into 6 equal-sized epochs of 20.5 seconds length, which are considered as input samples to the system. Moreover, the reliability of EEG characteristics depends on the length [59], so that this length seems to be the best tradeoff between number of epochs and signal length. Thus, with two recording sessions, we obtained 12 epochs for each participant, and from each epoch we extracted 16 feature vectors.

We estimated the **power spectral density (PSD)** as the single-sided amplitude spectrum from the Fast Fourier Transform of the signal.

Furthermore, we estimated multivariate autoregressive coefficients, from which we extracted a set of 14 measures of interaction between all of the 27 selected sensors. Estimation was based on the multivariate autoregressive model (MVAR) [17], [60]:

$$Y(t) = \sum_{k=1}^{P} A(k)Y(t-k) + U(t) \tag{1}$$

where $Y(t) = [y_1(t), ..., y_M(t)]^T$ is a vector holding the values of the M channels at time t, P is the model order, $A(k)$ are $M \times M$ coefficient matrices in which the element $a_{ij}(k)$ describes the dependence of $y_i(t)$ on $y_j(t-k)$, and $U(n)$ is the innovation process for segments with n data samples, which is assumed to be composed of white and uncorrelated noise. We used the functions mvfreqz.m and mvar.m from the BioSig toolbox [61] with model order $P = 41$. The maximum model order was chosen in order to capture slow signal spreads, but limited by the need to obtain a large ratio $n/(M \cdot P)$. The latter is needed to get an accurate model estimation [53]. In this study, this resulted in a ratio of $20.5 \cdot 500/(27 \cdot 41) = 9.26$, which is well above an acceptable threshold. In contrast to previous work, the sampling rate of 500 seems to oversample the signal, and the model order is quite high. However, one original contribution of this work is also to extend previous research to the multivariate domain, as we are interested in time-shifts of signals across the examined sensors, in the sense of spreading activity - often interpreted as so-called directed networks of the brain.

In order to fit the multivariate autoregressive model, we used partial correlation estimation with unbiased covariance estimates [60], which was found to be the most accurate estimation method in a comparative study [53].

The matrices $A(k)$ of size $M \times M$ formed the second feature, representing the **multivariate autoregressive coefficients (AR)**.

The other 14 features were obtained as follows. The estimated MVAR model was transformed from the time-domain into the z-domain and the f-domain, which accordingly yields two transfer functions. The multivariate parameters in the frequency domain that can be derived from these transfer functions were computed for 1 Hz frequency steps between 1 and 125 Hz: direct causality (DC) [62], spectrum (S) [63], transfer function (hh) and transfer function polynomial (AF) [64], real valued coherence (COH) and complex-valued coherence (iCOH) [65], partial coherence (pCOH) [66], partial directed coherence (PDC) and partial directed coherence factor (PDCF), generalized partial directed coherence (GPDC) [67], directed transfer function (DTF) [68], direct directed transfer function (dDTF), full frequency directed transfer function (ffDTF) [69], and Geweke's Granger Causality (GGC) [70], [71].

Next, PSD and all frequency-dependent measures of interaction, that is, all but DC, were averaged in classical frequency ranges delta (δ, 2-4 Hz), theta (θ, 5-7 Hz), alpha (α, 8-13 Hz), beta (β, 14-30 Hz), and gamma (γ, 31-80 Hz).

All frequency dependent measures were analysed once with this 5-band frequency configuration, and once by restricting to the 3 frequency ranges θ, α, and β, since this range has been shown to be more informative by Maiorana et al. [40]. The non-frequency dependent measures DC and AR were calculated on the band-pass filtered data (5-30 Hz) for this purpose.

B. Feature fusion

The autoregressive coefficients were obtained as 27×27 matrices for each $k = 1...P$. We concatenated these values as one long feature vector consisting of all $27 \times 27 \times 41$ coefficients. For PSD, we concatenated the values from all 27 electrodes from all 5 frequency bands, thus resulting in 27×5 values in the feature vector. For each of the 14 measures derived from the autoregressive model, we obtained interaction matrices of size 27×27, thus one value for each electrode combination. All measures but DC were frequency specific. Frequency specific interaction matrices were available separately for each of the 5 frequency ranges. Depending on whether the measure was directed or not, this matrix was symmetric (not-directed measures, e.g. coherence) and therefore containing redundant elements or not symmetric (directed measures, e.g. directed transfer function). We concatenated all non-redundant values from these interaction matrices for all frequencies of interest. For non-directed measures we took the upper triangular of the interaction matrix and concatenated these values for each frequency range. For measures without time-lagged auto-correlation, the diagonal of the interaction matrix was excluded because it contained no information. This resulted in high-dimensional feature vectors v of lengths ranging from $27 \times 27 = 729$ (DC as the only measure without frequency dimension) to $27 \times 27 \times 5 = 3645$ (all directed measures with autocorrelation).

C. Cross validation: enrolment and authentication

Enrolment consisted in creation of a template. The template consisted of the average of the feature vectors v of several epochs $e_1...e_k$ that belonged to one subject p. Thus, one element i in the template feature vector $\mathcal{V}$ of subject p was obtained as

$$\mathcal{V}_p(i) = 1/k \sum_{l=1}^{k} v_{p,e_l}(i) \tag{2}$$

Note that these epochs do not necessarily need to be all of the epochs that belong to one subject. For cross validation it is essential to separate enrolment and authentication. The comparison score Z is then obtained as the Pearson correlation ρ of the feature vector v of one subject's epoch, that is, epoch e_l of subject p_a, with some $\mathcal{V}_{p_b}$:

$$Z = \rho(v_{p_a,e_l}, \mathcal{V}_{p_b}) \tag{3}$$

In case of $a = b$, probe v and template $\mathcal{V}$ are mated. The Pearson correlation coefficient of a feature vector with a template builds an index of similarity between mated pairs and non-mated pairs. Correlation over the feature vectors allows adequate scaling across varying feature vector length.

A high correlation coefficient is close to 1 and indicates similarity, which is desirable when correlating epochs of one subject with the respective subject's template, that is, in the *mated* situation. A low correlation coefficient is close to 0 and indicates dissimilarity, which is desirable when correlating epochs of one subject with templates of other subjects, that is, in the *non-mated* situation.

D. Computation of performance metrics

The correlation was calculated between each probe (i.e., each feature vector v of each epoch e) and each template $\mathcal{V}$ (i.e., the template obtained during enrolment). The mated scores were the Pearson correlation coefficients between feature vectors of one subject's epoch with their respective template, as obtained according to the cross-validation procedure (see section II-C). This yields the mated scores $Z_{p_{aa},e}$ and non-mated comparison scores $Z_{p_{ab},e}$ as follows:

$$Z_{p_{aa},e_l} = \rho(v_{p_a,e_l}, \mathcal{V}_{p_a[!e_l]}) \tag{4}$$

$$Z_{p_{ab},e_l} = \rho(v_{p_a,e_l}, \mathcal{V}_{p_b}) \tag{5}$$

The index $[!e_l]$ in equation 4 means that the template did not include the epoch to be correlated with (see section II-C). The non-mated scores in equation 5 were the correlation coefficients obtained after Pearson correlation of feature vectors of one subject's epoch with all other subjects' templates, as obtained according to the cross-validation procedure (see section II-C).

Now, from the distribution of mated and non-mated scores, that is, correlation coefficients, one can find the intersection of the two empirical distributions and thus define the equal error rate (EER), that is, the error rate where the false acceptance rate equals the false rejection rate. Moreover, we calculated the ZeroFMR (the lowest false rejection rate for false acceptance rate =0) and the FMR1000 (the lowest false rejection rate for false acceptance rate≤0.1%). Approximate confidence intervals (95%) were calculated using nonparametric bootstrap and 1000 replicas.

In addition, we computed receiver operating characteristic (ROC) curve, the area under the curve (AUC) alongside with 95% confidence intervals on the true positive rate by threshold averaging and sampling using bootstrap and 1000 replicas.

E. Feature selection algorithm

The high-dimensional vectors are likely to contain redundant information, since neighbouring frequencies and neighbouring electrodes are likely to share information. We implemented a feature subset selection algorithm in a three-layered cross-validation procedure:

1) In the outer layer, we randomly partitioned the set of 60 subjects into 6 sets for a 6-fold cross-validation. Thus, in 6 iterations, each time 10 subjects were left out as the *test set*, the other 50 were submitted to the middle layer.
2) In the middle layer, the 50 subjects were again divided into 5 subsets, for a 5-fold cross-validation. Thus, in 5 iterations, each time 10 subjects were left out as the *evaluation set*, the other 40 were submitted to the inner layer as the *training set*.
3) In the inner layer, these 40 subjects formed the *training set* and were used in order to optimize the feature vector as follows:
   - The feature vector entries $i = 1...k$ were sorted by the smallest ratio σ_δ of within subject to between subject standard deviation σ

     $$\sigma_\delta(i) = \frac{\overline{\sigma_e(v_{p,.}(i))}}{\sigma_p(\mathcal{V}(i))} \tag{6}$$

     That is, the numerator was the mean over all subjects' p within subject standard deviation σ_e (i.e, the standard deviation across all epochs e), whereas the denominator was the between subject standard deviation σ_p (i.e., the standard deviation across all templates $\mathcal{V}$ of all subjects p). Please note that according to Section II-C, these were 6 epochs for the single-EEG scenario, 12 epochs for the pooled EEG-scenario, and 6 epochs for the cross-session-validation scenario.
   - The first 5 entries of the sorted feature vector, that is, those entries with the smallest value σ_δ initialized the subset of selected features.
   - Enrolment, authentication, and computation of error rates were done for this subset. Thus, for each subject p, an enrolment template was created as the average of all epochs e but one that was left out from the enrolment set. Then, this template and the left out epoch were used to create a mated score, while for all other subjects, the average of all epochs was calculated and correlated with the left out epoch of subject p, yielding non-mated scores. These values were used to compute the EER (see Sections II-C and II-D).
   - Then, we added each entry of the sorted feature vector stepwise and repeated the calculation of the EER as described in the previous step. Upon each added feature vector entry, we evaluated whether the resulting EER was larger or equal than in the previous step. If it was larger or equal, we excluded the entry from the feature subset, if it was reduced, we kept it. In order to save computing time, we implemented a stopping criterion for this iteration if there was no further improvement over 50 consecutive entries of the feature vector.
4) The optimized feature vector was tested with the remaining 10 subjects of the *evaluation set*, that is, those that were left out in the middle layer. Since this middle layer involved a 5-fold cross validation, this resulted in 5 optimized feature vectors. From these, the average length of the feature vector was calculated across these 5 vectors. The features from the 5 vectors were sorted according to how often they were selected in the 5 iterations. The final feature vector was filled from this sorted list, in order to be as long as the average of the 5 feature vectors, and in order to contain the top-most selected features.
5) This optimal feature vector was then used to calculate the mated and non-mated scores from the 10 subjects of the *test set*, that is, those that were left out in the outer layer. For reporting and statistics, the performance metrics EER, Zero-FMR, FMR1000, and AUC were calculated over all mated and non-mated scores from these 6 iterations. For each of the measures, we created histograms over the selected indices that built the six feature vectors resulting from cross validation. This allowed to assess the variation of the selection process across the 6 partitions.

Sorting the features initially by the standard deviations within subjects is a trivial but nevertheless efficient way to shorten the time-consuming step of a wrapper-style feature subset selection procedure. Please note that similar approaches implemented principal component analysis [72], [73], independent component analysis [74], or benchmarked the feature selection according to principal component analysis by the ratio of within-subject to between-subject variability [36].

F. Merging of multiple features

A recent review by Yang et al. [11] concluded that newer EEG features do not perform better than AR coefficients or PSD. Therefore, we merged the feature vectors of AR, AF, which is the frequency-dependent variant of AR, and pCOH into one long feature vector. Here, pCOH was selected as an improved version of the coherence, which was also used in several studies according to Table I and because our experimental results showed that the system performance is better with pCOH than with COH. The resulting vector was submitted to the same feature-subset selection procedure as the single-feature vectors, as described in section II-E. Also for this feature vector we performed an additional analysis when restricting the frequency range to $\theta - \beta$.

The applied merging technique of concatenating the feature vectors is also known as merging on feature-level.

G. Assessment of moderators

Because biometric systems should be robust against human factors, we performed the calculation of EER, Zero-FMR, FMR1000, and AUC with confidence intervals separately for subgroups of female and male participants. Moreover, we divided the sample into 3 age groups:

- $age < 36$, since most studies have used samples in this age range;
- $35 < age < 63$, in order to match the extended range of another study [26]
- $62 < age$, in order to show results based on participants that represent an older sample than what all but one of the previous studies [27] included as a dataset

The oldest group was not older than the oldest sample used in previous work, that is above 76 years, because this is also the maximum age in our subsample. In addition, since the data set used in this study was collected in a larger study involving clinical populations, we assessed whether the pathological changes in the EEG, which are prominent in patients with temporal lobe epilepsy, might affect the result. Therefore, we analysed the following subpopulations. Healthy controls (HC), patients with temporal lobe epilepsy (TLE), and patients with mild cognitive impairment or subjective cognitive complaints (MCI/SCC). Please note that we merged TLE with a focus on the left and right hemisphere into one group as well as patients with MCI and SCC into one group, in order to obtain samples that were large enough and still homogeneous in terms of the pathological changes that could be expected.

For the purpose of sub-group analyses, we calculated AUC with confidence intervals based on the mated and non-mated scores of the subjects from the respective subgroups that were obtained with the feature that was the best one so far. We did not include all of the available impostor scores for this calculation but only the within sub-group impostor scores, because we assumed high within-group similarity. Including all impostor scores would therefore probably overestimate the performance, since we consider within group matches, but between-group impostors. In contrast, the way we employed the calculation of the performance metrics allowed to estimate the performance when the groups were homegeneous in terms of the assessed factors.

Nevertheless, the subgroups were not homogeneous in terms of other factors, so that estimation of interactions between the effects of sex, age, and pathology was not possible. Therefore, in order to statistically characterize the effects and interactions of the between-subjects factors sex (female, male), age (young, middle, old), and pathological group (TLE, SCC/MCI, HC), as well as the within-subjects-factors cross-validation (single, pooled, and cross-session), and frequency restriction (whole range vs. restriction to $\theta-\beta$), we calculated a semi-parametric MANOVA, because this analysis method was designed in order to handle small sample sizes. This calculation was performed on mated scores, since we did not want to base it on some specific metric or threshold such as the EER or AUC. The non-parametric MANOVA requires metric data, but allows for non-normality and variance heterogeneity [75]. This method is implemented in the R-package MANOVA.RM [76]. We used it with the parametric bootstrap (1000 iterations) which showed the most favourable performance in unbalanced designs and was therefore generally recommended [75]. The Wald-Type statistic allowed us to address our multivariate hypotheses on between- and within-subjects factors and their interactions.

TABLE II
SAMPLE OVERVIEW

group	N	med. age	range	women	r-handed
MCI/SCC	27	64	48-76	13	26
TLE	13	48	21-66	9	12
HC	20	61.5	23-74	14	18

N= number; MCI= mild cognitive impairment; SCC= subjective cognitive complaints; TLE= temporal lobe epilepsy; l= left; r= right; HC= controls; med= median;

All mated scores of all single epochs were submitted to this statistic, but the model was informed about the participant to which the several epochs belonged. Thus the epochs were an additional factor that was not modelled. However, since the pooled-session variant included 12 epochs while the other two included 6 epochs, only 3 epochs from the first session and 3 epochs from the second session were considered for this statistic. To this end, we selected the first, third and fifth epoch from both sessions.

H. Experimental data

1) Sample: We recruited a total sample of 70 participants at the Department of Neurology, Paracelsus Medical University Salzburg, Austria, from May 2012 to December 2015 within a larger study focused on memory disorders. After exclusion of participants who did not undergo both EEG-examinations (two TLEr, one TLEl, three HC) or whose EEG was of poor quality (one SCC, one TLEl, two HC) 60 participants remained for this analysis. Poor quality of the EEG was defined as less than 4 sec in at least one of the two recordings after excluding segments of 500 ms according to the automatic data inspection (see Section II-H3). Table II gives an overview of the demographic characteristics of patients included in the subgroups.

We obtained 8 participants that were younger than 36 years, 27 participants that were up to 62 years old, and 27 participants that were aged 63 years or older.

2) Data registration: EEG was recorded in two sessions separated by two weeks and took place in the same setting, that is, in a quiet room. Participants were instructed to close their eyes and stay awake. Eyes closed is a condition that allows to reduce artefacts from blinking, and thus variability of the recorded EEG. Recordings lasted for 2-3 min. We used a BrainCap with a 10-20 system and a BrainAmp 16-bit ADC amplifier (Brain Products GmbH, Germany). The sampling rate was 500 Hz. Of the 32 recorded channels, one was used to monitor the lower vertical electrooculogram, and one was used to measure electrocardiographic activity. Two were positioned at the earlobes for re-referencing, which was conducted in order to remove the bias of the original reference, which was placed at FCz. Data analysis was conducted for data collected from the remaining 27 electrodes F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, Pz, FC1, FC2, CP1, CP2, FC5, FC6, CP5, CP6, TP9, and TP10. Impedances were kept below 10 kΩ.

The two EEG sessions were arranged to take place at the same time of the day. For most participants, this requirement was met by performing EEG within the same time-range around noon (1 pm). We aimed to keep the time difference between the two recordings below three hours. For three participants (HC, SCC, TLEl) the time difference was approximately four hours, for two patients (MCI, TLEr) the time difference was six hours, and for one HC the time difference was 11 hours.

3) Data preparation: Data was pre-processed with Brain Vision Analyzer (Version 1.05.0005, Brain Products GmbH). In order to re-reference all channels, a new reference was built by averaging the signal of earlobe electrodes. Butterworth Zero Phase Filters were used for a high-pass filter from 1 Hz (time constant 0.1592 s, 48 dB/oct), and an additional notch filter (50 Hz) was applied.

An automatic artefact detection was carried out in order to exclude highly contaminated datasets. Please note that the automation of this procedure ensures objectivity, which means at the same time that it is reproducible. Nevertheless, the nature and number of artefacts surely depends on the specific recording and participant. Maximal allowed voltage step per sampling point was 50 μV (values which exceeded this threshold were marked within a range of ±100 ms); maximal allowed absolute difference on an interval of 200 ms was 200 μV and lowest allowed absolute difference during an interval of 100 ms was 0.5 μV (values which exceeded this were marked with a surrounding of ±500 ms). The result of this artefact detection was reviewed visually in order to determine whether the automated detection yielded reasonable results and whether poor data quality was due to noise on the reference electrodes, which led to exclusion of the dataset.

The preprocessed data was exported into a generic data format and imported to Matlab® (release R2016b, The Mathworks, Massachusetts, USA).

III. Results

A. Results across features

Figure 1 shows large variation between the 16 assessed features. Nevertheless, for each of these features, the overestimation of performance of the biometric system is remarkable when enrolment and authentication were based on a single EEG session. Merging two EEG sessions yielded increased values for most measures, but when performing a hard cross-validation with strict division of enrolment and authentication into two EEG sessions, the error rates rose considerably. In the cross-session scenario, the feature AF yielded lowest EER, lowest Zero-FMR, lowest FMR1000, and largest AUC in comparison to all other features.

Figure 2 shows the results when restricting the frequency range to $\theta - \beta$, as done in Maiorana et al. [40]. In our sample this restriction had a negative effect on the performance metrics. However, AF still showed the best performance in comparison to all other features in the cross-session cross-validation.

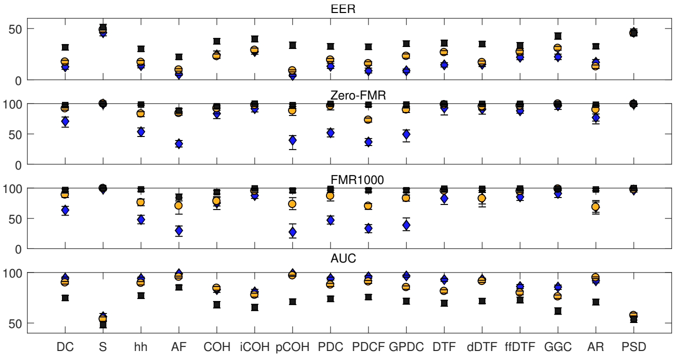

Fig. 1. EER, Zero-FMR, and FMR1000 separately for measures of interaction and three variants of cross validation: based on single EEG (blue diamond), pooled EEGs (yellow circle), and the cross-session scenario (black square).

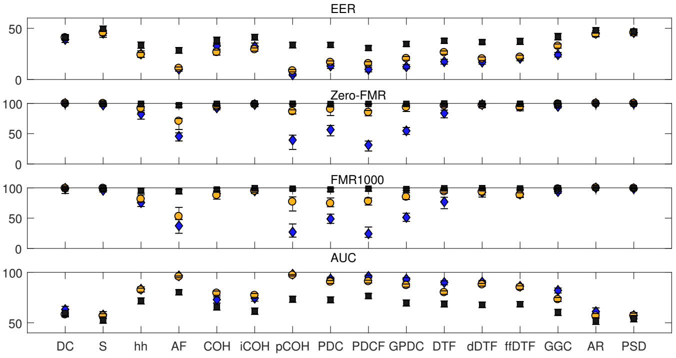

Fig. 2. EER, Zero-FMR, and FMR1000 separately for measures of interaction and three variants of cross validation when restricting the frequency range to $\theta - \beta$: based on single EEG (blue diamond), pooled EEGs (yellow circle), and the cross-session scenario (black square).

The best result was obtained with feature AF in the broad frequency band. Figure 3 shows the ROC curves for this feature. The single- but also pooled session scenario ROCs represented a clear overstatement of the actual performance, as denoted by the cross-session cross-validation scenario. In this scenario, the full frequency range was especially important for obtaining a low false positive rate.

B. Variation of feature selection

Figure 4 shows the histograms for all measures and the three types of cross validation for the analysis including all 5 frequency bands. The feature vectors represent the concatenated entries of the interaction matrices for each frequency. Thus, the lowest frequency starts with index 0, iterating over all electrode × electrode interactions, followed by the next frequency range and so on. Two aspects should be considered when evaluating this figure. First, the more similar the three histograms of single, pooled, and cross-session cross validation are, the more reliable is the respective feature. Second, vertical lines touching the upper border indicate that the respective feature was selected in all 6 cross-validation runs and, vice-versa, lower vertical lines indicate that the respective feature

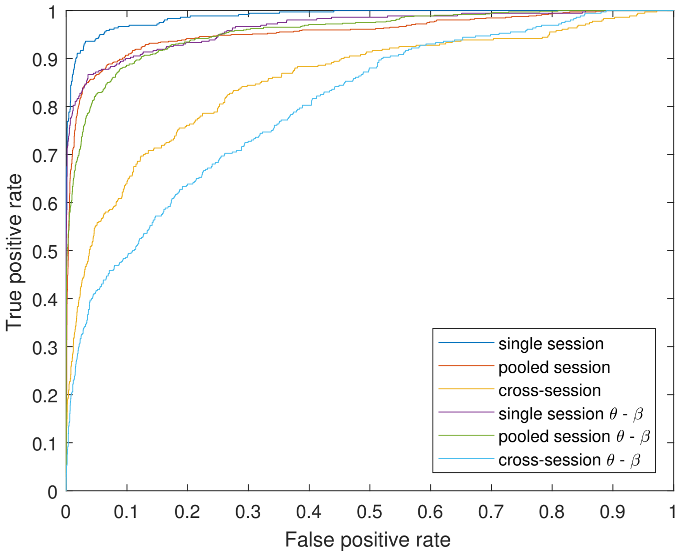

Fig. 3. ROC curves for feature AF for the broad-band $\delta-\gamma$ and the restricted-band $\theta-\beta$ variants.

was selected in a minor portion of cross-validation runs. Thus, the higher the vertical lines, the higher the reliability of the respective feature. For features that also showed low EERs, such as AF, many features were selected consistently in all cross runs, and this was also quite consistent over the three cross validation scenarios. For other features, such as S and PSD, most of the selected features were selected only in a subset of the cross-validation runs, suggesting that the feature selection process was affected by noise. For AR, the EER was low as well, but the feature vector varied a lot across the cross-validation runs and across the three evaluation scenarios.

C. Merging features

Table III shows the results for merging three features AF, pCOH, and AR for the broad-band $\delta-\gamma$ and the restricted-band $\theta-\beta$ variants in comparison to the single best feature AF. Also these results confirm the overestimation of performance of the biometric system when enrolment and authentication is based on a single EEG session or on two pooled EEG sessions. However, the single measure result was better than the merged result. This may be somewhat surprising, but is understandable on the background that we performed feature level fusion, while other fusion techniques on score or decision level [16] were reported to yield better performance. Moreover, the described feature subset selection method described in this work is very trustworthy, because it clearly separates the training, test, and evaluation sets, but it performed worse with increasing length of the feature vector. Thus, when the feature vector is too long, the most important information may not be detectable.

D. Moderators

Since AF resulted in the best performance, we performed subgroup analyses and statistical evaluations based on this method. Figures 5 and 6 show the AUCs and EERs with confidence intervals for sexes, age groups, and pathological subgroups separately for the full and restricted frequency range.

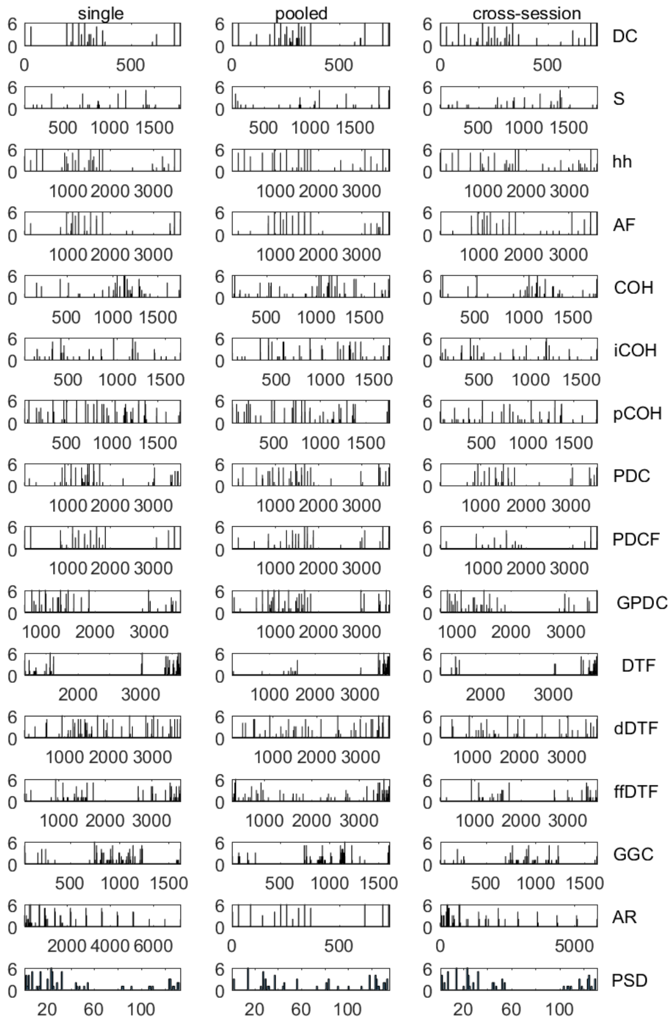

Fig. 4. Histograms for the selected features from the feature vectors. Rows represent the measures, columns the three scenarios of cross validation from left to right: based on one EEG, based on the pooled segments from both EEGs, based on cross-session cross validation. The maximum possible number of selections is 6, for the 6 cross-validation runs. Thus, if an entry on the feature vector yields a value of 6 this means that it was selected in all cross-session runs.

The difference between groups was negligible in the single session full frequency scenario, while there were notable differences in all other configurations. Because the cross-session scenario is the only one that has practical validity, we restrict the discussion to these results.

According to the AUC, the difference between female and male participants was negligible in the restricted frequency range scenario, but remarkable in the full frequency range, while EERs overlapped largely. In contrast to the expectations, the male subgroup yielded a lower AUC and a higher EER than the female subgroup.

The difference between pathological groups was most obvious in the full frequency configuration. Healthy controls showed the lowest AUC and highest EER, patients with

TABLE III
PERFORMANCE METRICS WHEN MERGING AF, PCOH, AND AR VS. AF ALONE

Scenario	EER	0FMR	FMR1000	AUC	AUC-CI
merged					
single	11.11	60.28	53.33	96.21	95.02-97.13
pooled	29.32	95.14	90.83	79.72	78.14-81.27
cross	29.46	99.72	96.67	75.82	73.1-78.06
single $\theta-\beta$	8.89	57.22	33.33	97.27	96.21-98.13
pooled $\theta-\beta$	11.11	76.94	63.06	95.51	94.56-96.26
cross $\theta-\beta$	28.64	98.33	95.28	80.56	78.08-82.58
AF					
single	5.34	34.44	32.50	98.85	98.25-99.23
pooled	9.70	84.44	68.47	95.51	94.42-96.50
cross	22.0	89.17	86.67	85.34	82.74-87.81
single $\theta-\beta$	10	47.22	41.39	96.47	95.20-97.43
pooled $\theta-\beta$	10.99	72.22	52.08	95.53	94.60-96.31
cross $\theta-\beta$	28.87	79.50	94.72	80.29	77.49-82.67

EER=Equal Error Rate; 0FMR=min. false rejection rate for false acceptance rate=0; FMR1000=min. false rejection rate for false acceptance rate≤0.1%; AUC= area under the curve; CI: confidence interval of AUC

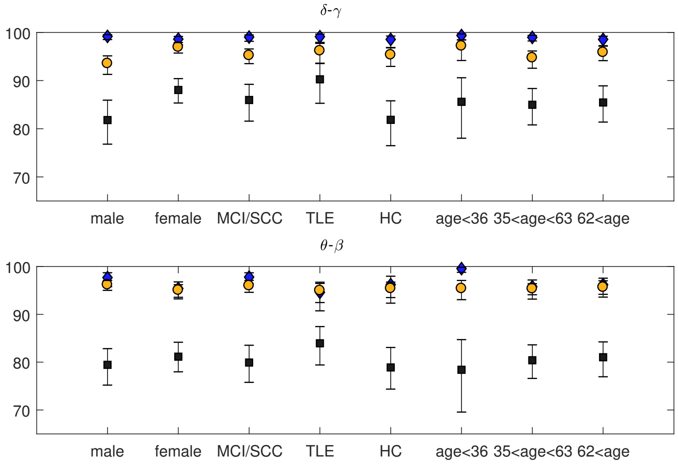

Fig. 5. AUC with confidence intervals (CI) separately for sexes, age groups, and pathological subgroups, based on measure AF for full and restricted frequency range.

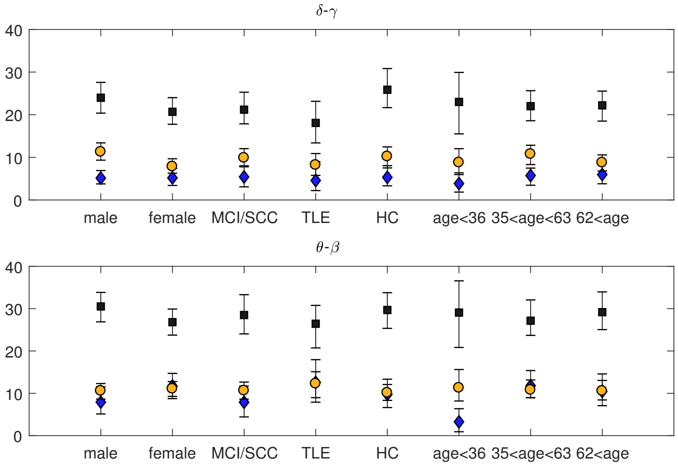

Fig. 6. EER with confidence intervals (CI) separately for sexes, age groups, and pathological subgroups, based on measure AF for full and restricted frequency range.

TABLE IV
RESULTS OF THE TWO SEPARATE MANOVAS FOR MATED SCORES BASED ON FEATURE AF WITH THE FACTORS SEX, AGE, SESSION, AND FREQUENCY (TOP), AND THE FACTORS SEX, PATHOLOGY, SESSION, AND FREQUENCY (BOTTOM), BASED ON THE WALD-TYPE STATISTIC (WTS).

factor(s)	WTS	df	p-value
sex	**10.45**	**1**	**.001**
age	2.21	2	.33
sex:age	0.55	2	.76
session	**511.80**	**2**	**<.001**
sex:session	**14.83**	**2**	**.001**
age:session	11.16	4	.03
sex:age:session	7.65	4	.11
frequency	1.12	1	.29
sex:frequency	**11.98**	**1**	**.001**
age:frequency	**7.66**	**2**	**.02**
sex:age:frequency	2.54	2	.28
session:frequency	0.85	2	.65
sex:session:frequency	**26.98**	**2**	**<.001**
age:session:frequency	**23.44**	**4**	**<.001**
sex:age:session:frequency	0.61	4	.96
sex	3.83	1	.05
patho	2.06	2	.36
sex:patho	**12.74**	**2**	**.002**
session	**547.26**	**2**	**<.001**
sex:session	**10.90**	**2**	**.004**
patho:session	**14.63**	**4**	**.006**
sex:patho:session	4.51	4	.34
frequency	0.01	1	.94
sex:frequency	**14.24**	**1**	**<.001**
patho:frequency	5.77	2	.06
sex:patho:frequency	6.63	2	.04
session:frequency	**8.86**	**2**	**.01**
sex:session:frequency	**16.29**	**2**	**<.001**
patho:session:frequency	**15.51**	**4**	**.004**
sex:patho:session:frequency	**26.30**	**4**	**<.001**

TLE highest AUC and lowest EER values, and patients with MCI/SCC were in between the other groups. Caution must be taken when interpreting the results of the MCI/SCC group, since this subsample does not contain any participant aged younger than 36, which is indeed an inherent property of this patient group [77]. It is therefore impossible to obtain a group of MCI/SCC patients that is directly comparable to the younger subjects in the healthy control group and the TLE group. Thus, the effect could be confounded by the relative age of this group.

The difference between the age groups in Figures 5 and 6 does not appear to be significant, since the confidence intervals overlap completely. However, the confidence interval for the youngest subgroup is larger for both the AUC and the EER.

The statistical analysis of the mated scores is given in Table IV. The Wald-Type statistic is larger with a larger effect or interaction of effects. p-values below .025 are considered to yield a significant effect or interaction after Bonferroni correction (represented in bold font). The interactions between sex and session as well as pathology and session suggest that the handling of multiple sessions was significantly affected by the sex and neurological conditions of the subjects in the dataset. Moreover, the interaction between age and session was marginally significant.

IV. DISCUSSION

In this study, we demonstrated the effect of several individual factors in interaction with the type of cross-validation, that is, whether it was based on one or two EEG sessions. As expected, the effect of cross-validation is large, and it interacts with individual factors. Our results emphasize the need of balancing the test samples in terms of age and sex and that it should be standard to report these aspects in biometric studies, in order to keep results from different studies comparable.

Moreover, by extending previous work on univariate autoregressive models, we could show that classical measures such as the PSD as well as advanced features extracted from the MVAR model and the respective coefficients are differentially affected by the handling of multiple EEG sessions. The finding that one of the derived measures yields better performance than the MVAR coefficients warrants further investigations.

A. Multivariate autoregressive modelling

We could show that the feature AF, thus, the frequency transform of a polynomial describing the transfer function of the multivariate autoregressive model, is slightly more informative than the corresponding MVAR coefficients, and that this feature is also quite reliable in contrast to most of the other examined features. However, merging this feature with other features did not increase, but decrease performance.

We may speculate that performance did not increase because AF combines both the advantages of the PSD as well as the AR. In the work of Maiorana et al. [40] the univariate AR was tested in a broadband variant and when restricted to the $\theta-\beta$ frequency range. In contrast, the AF allows a polynomial description of the model in each of the frequency bands of interest. It seems also that restricting this feature to a narrow frequency range reduced its information content.

Another possible explanation may be the feature subset selection technique we have implemented; the longer the feature vector becomes, the worse is the performance. The selection process iterates over a limited number of top-ranked feature vector entries and optimizes the vector within this range. The order with which the features are processed may change the local optimum that might be reached with this procedure. We implemented sorting by ratio of within to between subject variance, which may overemphasise feature vector entries that contain similar information and may fail to identify feature vector entries that yield complementary information.

B. Individual factors

We found a larger variance of performance metrics for younger participants than for the two older groups, a better performance for females than for males, and for patients with temporal lobe epilepsy than patients with MCI or SCC and healthy controls. Moreover, the factors age, sex, and pathology interacted significantly with session, and thus, with stability of the assessed biometric features. Stability was lower for males than females, and it was lower for healthy controls and young participants. The best stability was obtained in patients with TLE.

The selection of a broad frequency range might be of special relevance to the older participants in our experimental data set, where the EEG background rhythm shifts from the alpha to a lower frequency range, and for the pathological subgroups, were specific brain regions may be identified by pathological slowing or abnormal high frequency activity [78]. The importance of these findings is emphasised by the demographical changes, encouraging further research in older samples with frequent neurological pathologies such as MCI.

Several previous publications did not list the distribution of sex and/or age in their sample (see Table I). The sample of Maiorana et al. [40] was of similar size as our study with N=50, but the participants were all below 36 years, thus representing the subsample in our analysis with the greatest variance. The two studies that involve participants above this age range have a smaller sample size of 20 and 9 participants, respectively and show results that are worse in comparison to the other listed studies.

It is known that the menstrual cycle has an effect on EEG rhythms [58]. Since the two EEG sessions were separated by two weeks, this period corresponds to half of a menstrual cycle. This design provides data from two different phases of menstrual cycles for the younger female participants. Nevertheless, the female subgroup showed better performance metrics than the male subgroup. This is an interesting finding that warrants further investigation of EEG-variability in males.

The fact that TLE patients are well distinguishable from each other may be explained by the individual pathological details of the disease. They offer a variety of distinguishing factors in the EEG, such as the frequency of epileptic spikes, the presence of other epileptic patterns such as paroxysms, and intermittent theta or delta activity, again, favouring a broad frequency range configuration.

C. Multiple EEG sessions

Although the majority of studies published in the field is based on single sessions, the cross-session scenario is the only one valid for practical evaluation.

Our results replicate the range of possible results when using one or more EEG sessions. However, also sample size must be considered. If only a small sample is available, it was suggested to demonstrate the scalability of the results by calculating results for sub-partitions [79]. Another aspect is the way subset selection in the feature vectors was performed. We transparently described the selection procedure and performed a three-fold cross-validation with separate training, evaluation, and test sets, but not all publications conform to this practice, as criticised also by Maiorana et al. [40].

Recent studies demonstrated that a perfect match can be obtained also when comparing data from different sessions [18], [44]. One of these studies [18] is based on a sample size of 9 subjects, where the classification results of 100% were reported for the best performing channel combination. It is not clear how these subsets of channels were pre-selected out of originally 54 recording sites. The reported results may be biased by overfitting to this rather small sample. Moreover, the accuracy depends on which of two subject-subsets of 25% vs. 75% was used as the training or testing set. A lower accuracy of up to -10% of accuracy was reported when 75% of the sample were used as the training set. We assume that a 100% accuracy would not be obtained with a larger sample size.

Likewise, in [44] the 100% matches were obtained by selecting the electrode site and condition out of 150 possibilities where the classification was perfect. The reduction of recording sites was done after this perfect match was obtained, and the lowest number of electrodes needed was again tuned to the test set.

Classification by event related potentials (ERPs) seems to be an attractive approach, but works only for EEG recorded with cognitive stimulation. ERPs from a combination of cognitive tasks yield impressive performance of a biometric system [44], but enrolment and authentication takes a long time, while resting EEG is much faster. For example, the CEREBRE protocol [44] took 1.5 hours. It is possible that this duration can be reduced by limiting the protocol to the most meaningful stimulus categories and to a lower number of trials. However, it is well known and the authors state themselves that a reduction of trials for obtaining ERPs results in a lower signal to noise ratio and therefore, a lower classification performance.

D. *Outlook to future work*

The results for AF may be seen as an example. It is most likely that the effect of age, sex, and pathology in interaction with session is even stronger for those features with greater variation in the feature subset selection across the three cross validation scenarios, which may be subject to further investigation.

We used a data set with EEG recordings from two different days. This is the minimum for cross-session cross-validation, but more sessions can result in better performance of the biometric system because enrolment can be based on multiple sessions. However, by excluding one session for testing, the cross-validation is still valid [11]. It is thus supposed that the use of multiple sessions for enrolment could aid the selection of robust features for the biometric system [22]. However, enrolment over multiple days is not a realistic scenario of application of such a system. Nevertheless, it was also shown that depending on which sessions are combined, the performance metrics vary a lot [40], [43]. Thus, the session-to-session variability needs to be addressed ideally in > 2 sessions in future work.

The AR presented here was considered in a multivariate sense, while previous publications considered its univariate version. Despite the diagonale of the estimated covariance matrix representing the univariate AR, it is possible that the multivariate nature and, thus, complexity of the estimation process affects the accuracy of the result. The effect of univariate vs. multivariate autoregressive modelling might be evaluated in future work.

We implicitly normalised the feature vectors by employing Pearson correlation, which scaled efficiently with larger feature vectors. This assumes implicitly that the underlying relation is linear. For any other type of association, this approach is not optimal and other normalisation techniques available for biometric systems [80] should be evaluated.

Multifeature systems are rather standard than extraordinary in the high-dimensional situation of EEG data [81]. Fusion of features has been evaluated on feature, score, and decision level [16], [82]. It is possible that fusion on these higher levels leads to different results with respect to the susceptibility of stability to interindividual differences.

In order to have a comparable setting to previous work, we did not exclude artefacts. However, artefacts in the EEG (e.g., based on muscle movement) have a severe impact on test-retest reliability of EEG characteristics derived from the multivariate autoregressive model [59].

Additional factors can be hypothesised to play a role in multi-session evaluation. For example, we controlled time of the day, but it would be important to assess the robustness of an EEG system to these factors. Also intake of caffeine and tobacco could be an issue in biometric systems, as well as alcohol and any neurologically active medication.

Moreover, future applications need to consider the possibility of using user authentication via EEG in mobile communication systems, such as smartphones [12], [83], [84]. It is not likely that multiple channels are available for such an application, so that biometric feature vectors based on measures of interaction need to be optimised for low channel numbers.

Finally, a publicly available dataset that allows for the analysis of age, sex, and pathology in relation to EEG biometric features would be highly warranted. Unfortunately, the ethical approval of this study did not allow to make the dataset publicly available.

V. Conclusion

In this work, we pointed the community towards a serious problem that is present in the majority of published articles on EEG biometric features. The stability of EEG biometric features depends on factors such as age, sex, and pathology. This result emphasises the need for adequate datasets that involve at least two sessions when a biometric system should be tested, and accurate reporting of the factors that moderate stability. Furthermore, we demonstrate that the previous results obtained with autoregressive coefficients can be improved by deriving frequency-dependent measures.

Acknowledgment

We want to thank Eugen Trinka, Julia Höfler, Judith Dobesberger, Giorgi Kuchukhidze, Markus Leitinger, Margaritha Kirschner, Elisabeth Schmid, Daniela Sinadinoska, Susanne Grinzinger, Martin Scherr, and Wolfgang Staffen for recruitment of patients, and all participants for their time.

References

[1] H. Stassen, “Computerized recognition of persons by EEG spectral patterns,” *Electroencephalography and Clinical Neurophysiology*, vol. 49, no. 1-2, pp. 190 – 194, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0013469480903685

[2] M. Poulos, M. Rangoussi, V. Chrissikopoulos, and A. Evangelou, “Person identification based on parametric processing of the EEG,” in *Electronics, Circuits and Systems, 1999. Proceedings of ICECS ’99. The 6th IEEE International Conference on*, vol. 1, 1999, pp. 283–286.

[3] I. Nakanishi, S. Baba, and C. Miyamoto, “EEG based biometric authentication using new spectral features,” in *2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Jan 2009, pp. 651–654.

[4] K. Brigham and B. V. K. V. Kumar, "Subject identification from electroencephalogram (EEG) signals during imagined speech," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sept 2010, pp. 1–8.

[5] P. Cserti, B. Végsö, G. Kozmann, Z. Nagy, F. D. V. Fallani, and F. Babiloni, "Methods to highlight consistency in repeated EEG recordings," *IFAC Proceedings Volumes, 8th IFAC Symposium on Biological and Medical Systems*, vol. 45, no. 18, pp. 23 – 27, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474667016320675

[6] J. Chuang, H. Nguyen, C. Wang, and B. Johnson, "I think, therefore i am: Usability and security of authentication using brainwaves," in *Financial Cryptography and Data Security: FC Workshops, USEC and WAHC 2013, Okinawa, Japan, Revised Selected Papers*, A. A. Adams, M. Brenner, and M. Smith, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–16. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-41320-9_1

[7] G. Al-Hudhud, M. Alzamel, E. Alattas, and A. Alwabil, "Using brain signals patterns for biometric identity verification systems," *Computers in Human Behavior*, vol. 31, pp. 224–229, 2014.

[8] M. DelPozo-Banos, J. Alonso, J. Ticay-Rivas, and C. Travieso, "Electroencephalogram subject identification: a review," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6537–6554, 2014.

[9] M. DelPozo-Banos, C. Travieso, C. Weidemann, and J. Alonso, "EEG biometric identification: a thorough exploration of the time-frequency domain," *J Neural Eng*, vol. 12, p. 056019, 2015.

[10] G. Bajwa and R. Dantu, "Neurokey: Towards a new paradigm of cancelable biometrics-based key generation using electroencephalograms," *Computers & Security*, vol. 62, pp. 95 – 113, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404816300669

[11] S. Yang and F. Deravi, "On the usability of electroencephalographic signals for biometric recognition: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 958–969, 2017.

[12] J. Klonovs, C. K. Petersen, H. Olesen, and A. Hammershoj, "ID proof on the go: Development of a mobile EEG-based biometric authentication system," *IEEE Vehicular Technology Magazine*, vol. 8, no. 1, pp. 81–89, March 2013.

[13] K. Mohanchandra, L. GM, P. Kambli, and V. Krishnamurthy, "Using brain waves as new biometric feature for authenticating a computer user in real-time," *International Journal of Biometrics and Bioinformatics (IJBB)*, vol. 7, pp. 49–57, 2013.

[14] O. Sporns, "Cerebral cartography and connectomics," *Philos Trans R Soc Lond B Biol Sci*, vol. 19, no. 370, p. 1668, 2015.

[15] G. Van Baal, E. De Geus, and D. Boomsma, "Genetic influences on EEG coherence in 5-year-old twins," *Behavior Genetics*, vol. 28, no. 1, pp. 9–19, 1998.

[16] M. Garau, M. Fraschini, L. Didaci, and G. L. Marcialis, "Experimental results on multi-modal fusion of eeg-based personal verification algorithms," *International Conference on Biometrics (ICB)*, pp. 1–6, June 2016.

[17] R. Greenblatt, M. Pflieger, and A. Ossadtchi, "Connectivity measures applied to human brain electrophysiological data," *Journal for Neuroscience Methods*, vol. 207, pp. 1–16, 2012.

[18] D. La Rocca, P. Campisi, and G. Scarano, "On the repeatability of EEG features in a biometric recognition framework using a resting state protocol," *Biosignals*, pp. 419–428, 2013.

[19] S. Yang and F. Deravi, "On the effectiveness of EEG signals as a source of biometric information," in *2012 Third International Conference on Emerging Security Technologies*, Sept 2012, pp. 49–52.

[20] C. He, X. Lv, and Z. J. Wang, "Hashing the mAR coefficients from EEG data for person authentication," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 1445–1448.

[21] G. Van Baal, D. Boomsma, and E. De Geus, "Longitudinal genetic analysis of EEG coherence in young twins," *Behavior Genetics*, vol. 31, no. 6, pp. 637–651, 2001.

[22] A. Riera, A. Soria-Frisch, M. Caparrini, C. Grau, and G. Ruffini, "Unobtrusive biometric system based on electroencephalogram analysis," *EURASIP Journal on Advances in Signal Processing*, p. 143728, 2008.

[23] B. C. Armstrong, M. V. Ruiz-Blondet, N. Khalifian, K. J. Kurtz, Z. Jin, and S. Laszlo, "Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics," *Neurocomputing*, vol. 166, pp. 59 – 67, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215004725

[24] R. B. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles', "The electroencephalogram as a biometric," in *Canadian Conference on Electrical and Computer Engineering 2001. Conference Proceedings (Cat. No.01TH8555)*, vol. 2, 2001, pp. 1363–1366 vol.2.

[25] S. Marcel and J. D. R. Millan, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 743–752, April 2007.

[26] M. Kostílek and J. Stastny, "EEG biometric identification: Repeatability and influence of movement-related EEG," in *2012 International Conference on Applied Electronics*, Sept 2012, pp. 147–150.

[27] M. Naepflin, M. Wildi, and J. Sarnthein, "Test-retest reliability of resting EEG spectra validates a statistical signature of persons," *Clinical Neurophysiology*, vol. 118, no. 11, pp. 2519 – 2524, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138824570700394X

[28] C. Gupta, R. Palaniappan, and S. Swaminathan, "On the analysis of various techniques for a novel brain biometric system," *Int. J. Med. Eng. Inf.*, vol. 1, pp. 266–273, 2008.

[29] S. Sun, "Multitask learning for EEG-based biometrics," in *2008 19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.

[30] A. Yazdani, A. Roodaki, S. H. Rezatofighi, K. Misaghian, and S. K. Setarehdan, "Fisher linear discriminant based person identification using visual evoked potentials," in *2008 9th International Conference on Signal Processing*, Oct 2008, pp. 1677–1680.

[31] P. Campisi, G. Scarano, F. Babiloni, F. D. Fallani, S. Colonnese, E. Maiorana, and L. Forastiere, "Brain waves based user recognition using the eyes closed resting conditions protocol," *IEEE International Workshop on Information Forensics and Security*, pp. 1–6, Nov 2011.

[32] D. L. Rocca, P. Campisi, and G. Scarano, "EEG biometrics for individual recognition in resting state with closed eyes," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, Sept 2012, pp. 1–12.

[33] D. L. Rocca, P. Campisi, and J. Solé-Casals, "EEG based user recognition using BUMP modelling," in *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*, Sept 2013, pp. 1–12.

[34] D. La Rocca, P. Campisi, B. Vegso, P. Cserti, G. Kozmann, F. Babiloni, and F. De Vico Fallani, "Human brain distinctiveness based on EEG spectral coherence connectivity," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2406–2412, 2014.

[35] M. Fraschini, A. Hillebrand, M. Demuru, L. Didaci, and G. Marcialis, "An EEG-based biometric system using eigenvector centrality in resting state brain networks," *IEEE Signal Processing Letters*, vol. 22, pp. 666–670, 2015.

[36] E. Maiorana, D. L. Rocca, and P. Campisi, "Eigenbrains and eigentensorbrains: Parsimonious bases for EEG biometrics," *Neurocomputing*, vol. 171, pp. 638 – 648, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215009790

[37] R. Palaniappan, "Two-stage biometric authentication method using thought activity brain waves," *International Journal of Neural Systems*, vol. 18, no. 1, pp. 59–66, 2008.

[38] M. Abdullah, K. Subari, J. Loong, and N. Ahmad, "Analysis of the EEG signal for a practical biometric system," *World Academy of Science, Engineering and Technology*, vol. 44, pp. 1133–1137, 2010.

[39] F. Su, H. Zhou, Z. Feng, and J. Ma, "A biometric-based covert warning system using EEG," in *Biometrics (ICB), 2012 5th IAPR International Conference on*, 2012, pp. 342–347.

[40] E. Maiorana, D. La Rocca, and P. Campisi, "On the permanence of EEG signals for biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 1, pp. 163–175, Jan 2016.

[41] M. Jobert, F. Wilson, G. Ruigt, M. Brunovski, L. Prichep, W. Drinkenburg, and The IPEG Pharmaco-EEG Guidelines Committee, "Guidelines for the recording and evaluation of pharmaco-EEG data in man: The international pharmaco-EEG society (IPEG)," *Neuropsychobiology*, vol. 66, pp. 201–20, 2012.

[42] Y. Wang and L. Najafizadeh, "On the invariance of eeg-based signatures of individuality with application in biometric identification," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 4559–4562.

[43] R. Das, E. Maiorana, , and P. Campisi, "EEG biometrics using visual stimuli: a longitudinal study," *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 341–345, 2016.

[44] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "Permanence of the cerebre brain biometric protocol," *Pattern Recognition Letters*, vol. 95, pp. 37 – 43, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865517301940

[45] P. Nguyen, D. Tran, X. Huang, and D. Sharma, "A proposed feature extraction method for EEG-based person identification," Int'l Conf. Artificial Intelligence (ICAI), 2012.

[46] Q. Gui, Z. Jin, and W. Xu, "Exploring eeg-based biometrics for user identification and authentication," in *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Dec 2014, pp. 1–6.

[47] R. Palaniappan and K. V. R. Ravi, "A new method to identify individuals using signals from the brain," in *2003 Proceedings of the joint Fourth International Conference on Information, Communications and Signal Processing, and the Fourth Pacific Rim Conference on Multimedia*, vol. 3, Dec 2003, pp. 1442–1445 vol.3.

[48] G. K. Singhal and P. RamKumar, "Person identification using evoked potentials and peak matching," in *2007 Biometrics Symposium*, Sept 2007, pp. 1–6.

[49] R. Palaniappan, "Electroencephalogram signals from imagined activities: A novel biometric identifier for a small population," in *Intelligent Data Engineering and Automated Learning*, E. Corchado, Ed., vol. 4224. Berlin Heidelberg: Springer-Verlag, 2006, pp. 604–611.

[50] R. Palaniappan and D. P. Mandic, "EEG based biometric framework for automatic identity verification," *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 49, no. 2, pp. 243–250, 2007. [Online]. Available: http:\/\/dx.doi.org\/10.1007\/s11265-007-0078-1

[51] X. Bao, J. Wang, and J. Hu, "Method of individual identification based on electroencephalogram analysis," in *2009 International Conference on New Trends in Information and Service Science*, June 2009, pp. 390–393.

[52] S.-K. Yeom, H.-I. Suk, and S.-W. Lee, "Person authentication from neural activity of face-specific visual self-representation," *Pattern Recognition*, vol. 46, no. 4, pp. 1159 – 1169, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312004578

[53] A. Schlögl, "A comparison of multivariate autoregressive estimators," *Signal Processing*, vol. 86, pp. 2426–2429, 2006.

[54] C. Panayiotopoulos, *Optimal use of the EEG in the diagnosis and management of epilepsies*. Oxfordshire, UK: Bladon Medical Publishing, 2005.

[55] D. Barnes and K. Yaffe, "The projected effect of risk factor reduction on alzheimer's disease prevalence," *Lancet Neurol*, vol. 10, pp. 819–28, 2011.

[56] P. Rossini, S. Rossi, C. Babiloni, and J. Polich, "Clinical neurophysiology of aging brain: From normal aging to neurodegeneration," *Prog Neurobiol*, vol. 83, pp. 375–400, 2007.

[57] Y. Wada, Y. Takizawa, Z. Jiang, and N. Yamaguchi, "Gender differences in quantitative EEG at rest and during photic stimulation in normal young adults," *Clin Electroencephalogr*, vol. 25, pp. 81–5, 1994.

[58] C. Brötzner, W. Klimesch, M. Doppelmayr, A. Zauner, and H. Kerschbaum, "Resting state alpha frequency is associated with menstrual cycle phase, estradiol and use of oral contraceptives," *Brain Research*, vol. 19, pp. 36–44, 2014.

[59] Y. Höller, A. Uhl, A. Bathke, A. Thomschewski, K. Butz, R. Nardone, J. Fell, and E. Trinka, "Reliability of EEG measures of interaction: a paradigm shift is needed to fight the reproducibility crisis," *Front Hum Neurosci*, 2017.

[60] S. Marple, *Digital Spectral analysis with applications*. Prentice Hall, 1987.

[61] A. Schlögl and C. Brunner, "BioSig: A free and open source software library for BCI research," *Computer*, vol. 41, pp. 44–50, 2008.

[62] M. Kaminski, M. Ding, W. Truccolo, and S. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biol Cybern*, vol. 85, pp. 145–57, 2001.

[63] V. Murthy, "Estimation of the cross-spectrum," *Ann Math Statist*, vol. 34, pp. 1012–21, 1963.

[64] M. Eichler, "On the evaluation of information flow in multivariate systems by the directed transfer function," *Biol Cybern*, vol. 94, pp. 469–82, 2006.

[65] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett, "Identifying true brain interaction from EEG data using the imaginary part of coherency," *Clin Neurophysiol*, vol. 115, pp. 2292–307, 2004.

[66] W. Gersch and G. Goddard, "Epileptic focus location: spectral analysis method." *Science*, vol. 169, pp. 701–2, 1970.

[67] L. Baccalá, D. Takahashi, and K. Sameshima, "Generalized partial directed coherence," in *Proceedings of the 15th International Conference on Digital Signal Processing (DSP); July 1-4, Wales, UK*, S. Sanei, J. Chambers, J. McWhirter, Y. Hicks, and A. Constantinides, Eds. New York: IEEE, 2007, pp. 162–6.

[68] M. Kaminskí and K. Blinowska, "A new method of the description of the information flow in the brain structures," *Biol Cybern*, vol. 65, pp. 203–210, 1991.

[69] A. Korzeniewska, M. Maczak, M. Kaminskí, K. Blinowska, and S. Kasicki, "Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method," *J Neurosci Methods*, vol. 125, pp. 195–207, 2003.

[70] J. Geweke, "Measures of conditional linear dependence and feedback between time series," *J Am Stat Assoc*, vol. 77, pp. 304–313, 1982.

[71] S. Bressler, C. Richter, Y. Chen, and M. Ding, "Cortical functional network organization from autoregressive modeling of local field potential oscillations," *Stat Med*, vol. 26, pp. 3875–85, 2007.

[72] K. V. R. Ravi and R. Palaniappan, "Leave-one-out authentication of persons using 40 hz EEG oscillations," in *EUROCON 2005 - The International Conference on "Computer as a Tool"*, vol. 2, Nov 2005, pp. 1386–1389.

[73] R. Palaniappan and K. Ravi, "Improving visual evoked potential feature classification for person recognition using PCA and normalization," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 726 – 733, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786550500320X

[74] P. Tangkraingkij, C. Lursinsap, S. Sanguansintukul, and T. Desudchit, "Selecting relevant EEG signal locations for personal identification problem using ICA and neural network," in *2009 Eighth IEEE/ACIS International Conference on Computer and Information Science*, June 2009, pp. 616–621.

[75] A. Bathke, S. Friedrich, F. Konietschke, M. Pauly, W. Staffen, N. Strobl, and Y. Höller, "Using EEG, SPECT, and multivariate resampling methods to differentiate between Alzheimer's and other cognitive impairments," arXiv preprint arXiv:1606.09004, 2016.

[76] S. Friedrich, F. Konietschke, and M. Pauly, "Manova.rm: A package for calculating test statistics and their resampling versions for heteroscedastic semi-parametric multivariate data or repeated measures designs," R-Package Version 0.1.1, https://CRAN.R-project.org/package=MANOVA.RM, 2017.

[77] R. C. Petersen, O. Lopez, M. J. Armstrong, T. S. Getchius, M. Ganguli, D. Gloss, G. S. Gronseth, D. Marson, T. Pringsheim, G. S. Day, M. Sager, J. Stevens, and A. Rae-Grant, "Practice guideline update summary: Mild cognitive impairment," *Neurology*, 2017. [Online]. Available: http://n.neurology.org/content/early/2017/12/27/WNL.0000000000004826

[78] B. Frauscher, F. Bartolomei, K. Kobayashi, J. Cimbalnik, M. A. van't Klooster, S. Rampp, H. Otsubo, Y. Höller, J. Y. Wu, E. Asano, J. J. Engel, P. Kahane, J. Jacobs, and J. Gotman, "High-frequency oscillations: The state of clinical research," *Epilepsia*, vol. 58, pp. 1316–1329, 2017.

[79] Q. Zhao, H. Peng, B. Hu, Q. Liu, L. Liu, Y. Qi, and L. Li, *Improving Individual Identification in Security Check with an EEG Based Biometric Solution*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 145–155. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15314-3_14

[80] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, pp. 2270–2285, 2005.

[81] H. Jian-Feng, "Multifeature biometric system based on EEG signals," *In Proceedings of the 2nd International Conference on Interaction Sciences*, pp. 1341–1345, 2009.

[82] H. A. Shedeed, "A new method for person identification in a biometric security system based on brain EEG signal processing," in *2011 World Congress on Information and Communication Technologies*, Dec 2011, pp. 1205–1210.

[83] B. Hu, Q. Liu, Q. Zhao, Y. Qi, and H. Peng, "A real-time electroencephalogram (EEG) based individual identification interface for mobile security in ubiquitous environment," in *2011 IEEE Asia-Pacific Services Computing Conference*, Dec 2011, pp. 436–441.

[84] M. T. Curran, J. k. Yang, N. Merrill, and J. Chuang, "Passthoughts authentication with low cost EarEEG," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 1979–1982.

Yvonne Höller obtained a doctoral degree in psychology (2010), a Dipl.-Eng. degree in computer science (2012), and is pursuing a PhD in computer sciences at the Paris Lodron University, all three in Salzburg (Austria). She was a Post-Doc at the Paracelsus Medical University (2008-2018) and was then appointed as a Professor in Psychology at the University of Akureyri, Iceland. Her current research interests include quantitative EEG analysis and thereof machine learning and feature-subset selection in high-dimensional feature space.

Arne Bathke became Assistant Professor of Statistics in 2001, Associate Professor in 2007, director of graduate studies, and director of the Applied Statistics Laboratory at the Department of Statistics, University of Kentucky (USA), 2010-2012. He became a Full Professor at the Department of Mathematics, Paris Lodron University Salzburg (Austria), where he is currently the dean of the faculty of natural sciences. His main areas of research are nonparametric statistics, multivariate methods, statistics for repeated measures, and longitudinal data.

Andreas Uhl became an Associate Professor in 2000 at the Department of Computer Sciences, Paris Lodron University Salzburg (Austria), was a guest professor for computer science at the Johannes Kepler University of Linz (Austria) and the Klagenfurt University (Austria), before he was promoted to a Full Professor in 2012. His main research topics are multimedia security, biometrics, medical image analysis and classification, multimedia data transfer and storage, numbertheoretical numerics, and random number generation.